

Technical Report: The Last Word (fall 2019)

Team: Alvin Lo, Anu Kriti Wadhwa, Avinash Damania, Sonali Kondapalli, Sriram Hariharan

Introduction

Endangered languages are languages that are at risk of dying out and becoming extinct. A language is classified as endangered when there are very few native speakers which have a decreasing trend over time.

Causes that may contribute to the endangerment of a language can be categorized into three main groups - physical, socio-economic and political. Physical factors such as natural disasters, violent wars, and genocide which are powerful or prevalent enough to wipe out populations of native language speakers.

Socio-economic reasons include urbanization, where youth are moving from native areas to cities and bringing up their families there instead. Resultantly, the following generation would have low incentives to learn their native languages which are not used in the urban areas in which they reside. Language endangerment is also caused by cultural and economic marginalization, as youth may once again favor learning more popular and arguably, more functional languages for purposes such as career advancement. Inter-marriage between people from different native group languages could contribute to language endangerment. This is because, for convenience in communication among the family, any children they may have may only learn one of the two languages their parents speak.

An example of political marginalization could be when a government of a large nation adopts a singular official language. This causes smaller areas to have a language shift towards the official language for functional purposes, endangering their native languages. Political repression happens when leadership within nations promotes having a singular national culture that is focussed around one language. This immediately limits opportunities for people in other language groups, forcing them to shift from their native language to the promoted one. These groups tend to be minorities in the nations and may be forced to resettle, or even break away from their families. Over time, this results in potential language endangerment as well.

Motivation

Similar to endangered species, there are arguments which claim that the extinction of languages is due to evolution, and the idea of 'survival of the fittest'. Languages that prevail today are the most utilized as they are functional in communication among large groups of people, increasing productivity, efficiency, and convenience. However, with the loss of some languages, we would be losing not only their semantics, such as phrases, expressions or grammatical rules but also the wealth of meaning they hold in understanding the culture in

which they are used. These languages provide cultural identity and communal belonging. Without them, there is loss of information, traditions, songs, anecdotes, which may be historical, of medicinal value, or anthropological significance.

There are institutions and nonprofit organizations around the world that are dedicated to saving these languages. There are also multiple efforts in documenting and recording as much as possible to ensure that language-specific records remain and can be understood in the unfortunate scenario that the language becomes extinct.

There are also efforts being made in language revitalization, where endangered languages are being preserved by being taught to children, encouraging them to become fluent in the tongue.

Lastly, technology contributes to these efforts as well. Media such as digital classrooms, podcasts, audio recordings, phone applications and computer programs made in these endangered languages promote their use.

There is little awareness of these dying languages outside of the relevant communities, such as researchers in the field, charities for the cause, or native speakers. The resources we found on the topic tend to be either overwhelming data sets which are inaccessible to most, or dense publications which are difficult to interpret for most unless they are familiar with the terminologies. With our website, we hope to create a resource to raise awareness on the issue and provide relevant information to users in a simple, easy-to-comprehend fashion. In addition to parsing the data sets and presenting them in a more concise manner, we also hope to provide additional resources for those who would like to learn more. These resources could be general information on the issue of endangered languages or language-specific.

Pariona, Amber. "Why Do We Need To Save Dying Languages?" WorldAtlas, 8 Aug. 2017, <https://www.worldatlas.com/articles/why-do-we-need-to-save-dying-languages.html>.

Models

The three models we focused on are the information about **endangered languages**, facts about those **countries**, and **charities** that deal with those countries/languages that you can get involved with.

Our first model is languages. Each language has data about it such as its name, countries, country codes, ISO639-3 codes, degree of endangerment, alternate names, and the number of speakers.

Our second model is countries. Each country has data on its name, region, development level, capital, longitude, and latitude.

Our third model is charities. Each charity has data on its name, address, affiliation, asset amount, classification, deductibility, foundation status, income amount, ntee classification, ntee code, and ntee type.

For **endangered languages**, we hope to gather information like the status of the language (e.g. Threatened, Dormant), Classification (e.g. Semitic, Afro-Asiatic) and the development (e.g. Grammar). AS an estimation, there will be 2,580 instances of this model.



Hindi

ISO: hin

Classification: Indo-European, Indo-Iranian, Indo-Aryan, Western Hindi, Hindustani

Mainly Spoken: India

Location: Widespread in north India: northern Bihar, Himachal Pradesh, Madhya Pradesh, Punjab, Rajasthan, Uttar Pradesh, and Uttarakhand states; Delhi.

Braille script . Devanagari script , primary usage. Newa script , no longer in use, historic usage.

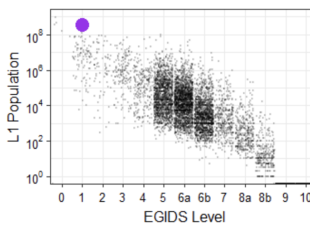
[LEARN MORE](#)

Hindi

Status: 1 . Statutory national language , also statutory provincial language in Bihar State and 12 other jurisdictions.

Mainly spoken in [India](#).

Speaker population: 612,000,000 in India, all users. L1 users: 339,000,000 in India . L2 users: 273,000,000 . Total users in Dialects: Khari Boli . Formal vocabulary borrowed from Sanskrit, de-Persianized, de-Arabicized. Literary Hindi, or Hindi-Ur India, refers here to the unofficial lingua franca of northwest India. Has a lexical mixture in varying proportions of Hindi a Language cloud:



[Charity to help!](#)

Each dialect is connected to the regions in which they are spoken and active, and the second model will present facts about these **countries**, such as the population, which languages are spoken there, the linguistic diversity, literacy, etc. This is estimated to have 195 instances.



India

Population: 1,266,884,000 (2017 World Factbook)

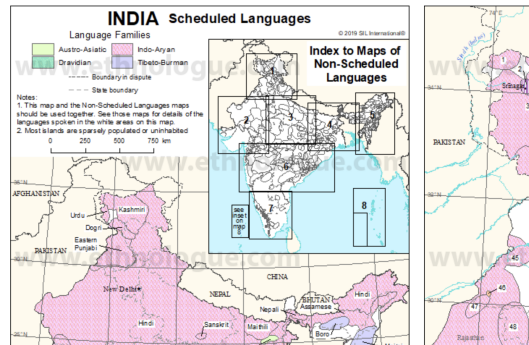
Principal Languages: English, Hindi

Language Counts: The number of individual languages listed for India is 460. Of these, 447 are living and 13 are extinct. 119 are institutional, 138 are vigorous, 112 are in trouble, and 14 are dying.

Literacy Rate: 71% (2017 World Factbook)

Languages: [Hindi](#)

[Charity to help!](#)



The third and last model will provide links to different **charities** and organizations that help preserve linguistic diversity. We include links to the donation pages, their website, as well as a display of which regions they're active in as well as which languages they will be serving.

LIVING TONGUES
 INSTITUTE FOR ENDANGERED LANGUAGES

Living Tongues

Institute for Endangered Languages

Active in:

Africa

Asia

Oceania

North America

Central and South America

Living Tongues Institute 4676 Commercial St
 SE, # 454 Salem, Oregon, OR 97302

[LEARN MORE](#)

Official Name: Living Tongues



[Donate Here](#)

[Website](#)

Languages Helping: [Irish](#) [Acheron](#) [Arapaho](#)

Countries Serving: [Ireland](#) [Burkina Faso](#) [Guatemala](#)

Mission: BRINGING VOICES TO THE FUTURE Assisting indigenous communities in their struggle for cultural and linguistic survival economically, or socio-culturally dominant ones. Every two weeks the last fluent speaker of a language passes on and with the ancestral tongues. Nearly half of the world's languages are likely to vanish in the next 100 years. The mission of the Living Tongues Institute is the preservation, and revitalization of endangered languages worldwide through linguist-aided, community-driven multimedia work with the last speakers of local endangered languages. After we obtain the permission of the community to work with them, we provide revitalization, etc. program. Story books, basic literacy materials as well as grammatical and lexical materials in electronic format and archive our video for the use of future generations. See our publications here. Community Training Involving indigenous reverse declining prestige, bridge the digital divide, and increase the range of uses of minority tongues. We train community documentation project to succeed and be embraced by the speech community, and creates a legacy for future generations. I

User stories

1. Add a non-default favicon tab, so it is easier to navigate to the page when multiple tabs are open.

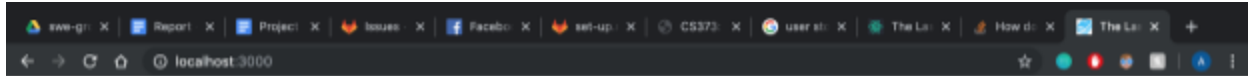




Figure 1: Favicon when multiple tabs are open

As Figure 1 shows above, we changed our favicon tab from the generic  to , to make it more easily identifiable by users who may have multiple tabs open.

2. Have a more attractive splash page

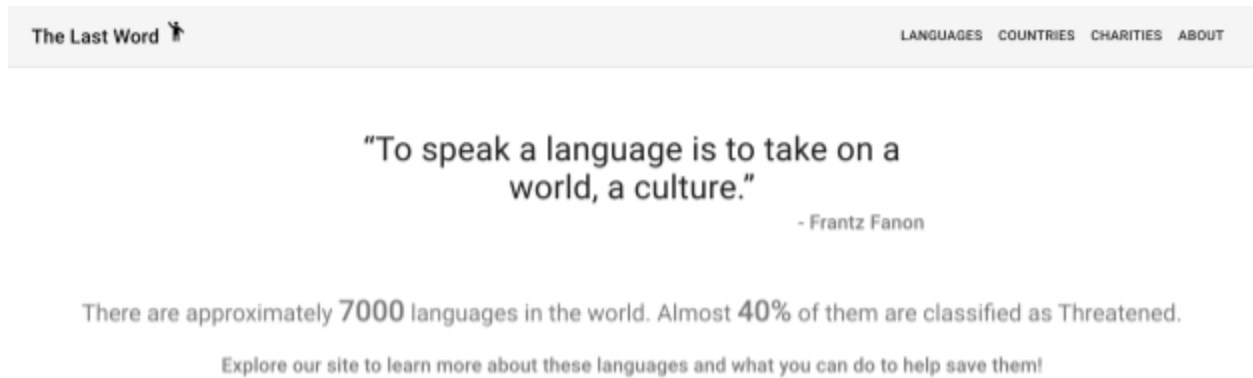


Figure 2a: Previous Splash Page



“To speak a **language**
is to
take on a **world**, a
culture.”

- Frantz Fanon

Almost 40% of the languages in the world are
classified as “threatened”.

If these languages were to become extinct,

© Copyright 2019 SWEC Group Ltd

Figure 2b: Updated Splash Page

We updated the splash page to make it more attractive, by adding more color, images, and emphasis on different portions. Figure 2a is what we had previously, while Figure 2b reflects these changes.

3. We added information for the main spoken languages in each country, as in Figure 3.

Canada

Main Languages: English, French

Figure 3: Main Languages of Country

This change can be seen on the ‘countries’ page.

4. Cards on about page not clickable

We did make them not clickable when we were still using media cards. However, we updated the page to have a tabular format, and this user story no longer applied to our website layout. This is seen in Figure 4 below:

The screenshot shows a web browser window with the URL `localhost:3000/languageInfo`. The page has a blue header with the logo 'The Last Word' and navigation links: 'LANGUAGES', 'COUNTRIES', 'CHARITIES', and 'ABOUT'. Below the header, the 'Languages' section is active, featuring a search bar. The main content is a table with the following data:

Name	Alternate Names	Countries	Endangerment	Speakers
South Italian	Neapolitan; Neapolitan-Calabrese; неаполитанский; неаполитанско-калабр...	Italy	Vulnerable	7500000
Sicilian	None	Italy	Vulnerable	5000000
Low Saxon	Low German, Niedersächsisch, Nedersaksisch, Niederdeutsch, Plattdeutsc...	Germany, Denmark, Netherlands, Poland, Russian Federation	Vulnerable	4800000
Belarusian	None	Belarus, Latvia, Lithuania, Poland, Russian Federation, Ukra...	Vulnerable	4000000

At the bottom of the page, there is a copyright notice: '© Copyright 2019 SWE Group 18'.

Figure 4: Tabular format

Sorting/Searching/Filtering:

For our sorting/searching/filtering implementation, we used a react library, called `material-table`, which already had sorting, and searching implemented within it's table component. We gave the Table component access to all the data (in the form of a JSON array) that is fetched using our back-end API, with specific endpoints to fetch "all" the data for a specific model type. We then provided different "functions" and exposed pieces of data which made the searching and sorting of the data provided much easier. We have the sorting, searching, and filtering working for all 3 of our model types. For each Model, you are able to sort, search, and filter the following fields:

Languages: Name, Spanish Name, Countries, Severity, Population

Countries: 3 Letter Code, Name, Region, Development, Capital

Charities: Name, Ein, Classification, Income Amount, Income Assets

Currently, we are able to handle partial matches (due to material-table being able to handle it), but are not able to handle multiple terms.

Tools

Material table

We used a react library, called material-table, to display the sets of data for our models. We chose this library primarily because it worked with our existing react code. It is fully customizable, which made it convenient to reformat to fit the data sets we needed. Additionally, it allows for easy navigation across each table, provides tools to manipulate and query the data, and lastly would allow for a corresponding visualization for future parts of the project.

We implemented the **search** using the properties of the material table library as well. We chose to do this as it was the most convenient method to implement search with the existing code we had, in a manner that allowed it to be customizable should we change the representation of our data in the future.

Cheerio/Beautiful Soup

Initially, we used Beautiful Soup to scrape Ethnologue. However, in the past two weeks, they updated their paywall and we were no longer able to scrape from the site. In light of this, we looked for new sources of data for endangered languages which are the following:

The new country API:

1. <https://datahelpdesk.worldbank.org/knowledgebase/articles/898590-country-api-queries?fbclid=IwAR2lu-121C1yJ9cCDFzslq5CZfno0XLHa-TBbFH1phfXAIJhV9HFxeiAJ3o>
2. http://api.worldbank.org/v2/country?format=json&fbclid=IwAR2KLsh6BxpEP0vKMh8Hep7lWiQ1i1Do2FaPHOu4F_wY7RrU0wtbnnnt-rU

The new language API:

1. https://www.theguardian.com/news/datablog/2011/apr/15/language-extinct-endangered?fbclid=IwAR1MBCFP8xFgqloPbs3mmY-1fPGK2YshCcfozgDFiH_FuKkQ4gLYf1KDDPg

For our Charities, we did a scrape for charities related to the keyword “language” on a charity navigator and scraped all the data for them.

We used cheerio and beautiful soup to scrape these sites for the information we needed. They

Amazon Relational Database Service (RDS)

We chose to use Amazon RDS as it made our backend database easy to set up, operate and scale for possible future needs, all in the cloud. We used only the free tier of the service.

RESTful API

For the RESTful API, we scraped the website 'ethnologue.com' for the raw language information. This includes the status of the dialect, which allows the language to be classified as endangered. We also used REST Countries, a very popular API for collecting information about various nations. This we used this to gather statistics about the countries from which the endangered languages originate. The last API we implemented was the Pixabay API to lend stock photos, a visual component, to the pages of information that we offer. Since it is difficult to gather photos for every obscure language, Pixabay makes that easy by providing us relevant pictures for the website.

We wrote some tests for our backend and API using Postman. For the backend, we used the inbuilt unit test library, and basically just tested the manipulation of the data once we got it into the forms we want, and the scraping of the data from cached HTML pages so we don't need to test with live connections. We also tested some utility functions that we had used all around the app. For the Postman side of things, we just used the Postman desktop app and Collections tester to create some unit tests using Postman's unique version of js. The output and log data can be seen. They are also on the API documentation.

Node is designed to build scalable network **Tools**

Beautiful Soup

To scrape from the Ethnologue website, we used the BeautifulSoup Python package. We scraped only selected information which includes:

- Language data representation, given the language abbreviation as defined by ISO 639-2
- All the data we can find about all the languages in a country, given the country code
 - Country's main page
 - Language status profile of for languages in that country
 - Maps of the languages for each country
- The map links and titles from the map page, given the page on Ethnologue
- The languages in a country, given the page on Ethnologue
- Key information, such as title and main body, given the main page for a country
- Language status profile
- Text from a language page, such as language name, and its data

Flask

Flask is a lightweight Web Server Gateway Interface (WSGI) web application framework. WSGI is a specification that describes how a web server communicates with web applications, and how web applications can be chained together to process one request.

We decided to use flask because of the convenience to scale up our application at a later time. It is also a flexible solution, not enforcing any dependencies or project layout, allowing us to shape our website the way we wish to.

React JS

React is a JavaScript library for building user interfaces. We decided to use React from the get-go to make our future work regarding the dynamism of the website easier. It enables us to create an interactive user interface moving forward. It is also an elegant solution for the back-end of the website, as it is simple to debug as compared to other alternatives. We used create-react-app to bootstrap our application and added code on top of that to obtain our web app.

We used the 'material-UI' and 'react-router' frameworks for the front end, which are both React libraries. Material Design is a unified system that combines theory, resources, and tools for crafting digital experiences. (Woodhead) We used 'material-UI' as we wanted to follow a 'best practices' approach to creating our website, by keeping material design in mind from the very start. React router, on the other hand, allows us to build a single-page web application with navigation without the page refreshing as the user navigates. Again, keeping the user in mind while designing, this will enable the user to have a smoother experience while perusing our site.

The Documentation for our API/backend can be found at:

<https://web.postman.co/collections/4350123-9a28d01f-8e60-46c3-b030-0e2f6704d480?version=latest&workspace=0db02365-69e7-4923-8711-dc48241a01e6>

The backend is hosted at:

<http://the-last-word-backend-254800.appspot.com/>

We provide 3 endpoints currently:

GET META

 Comments (0)

<https://the-last-word-backend-254800.appspot.com/meta>

Returns a JSON array of data on the contributors to this project (pulled from the gitlab repo)

GET LANGUAGE_DATA

Comments (0)

```
https://the-last-word-backend-254800.appspot.com/language/:iso_code
```

by passing in the three-letter ISO 639-2 Code for a language to this endpoint, given a JSON object with data on the language, as well as urls of pictures and graphs relating to that language.

Path Variables

iso_code	eng
----------	-----

GET COUNTRY_DATA

Comments (0)

```
https://the-last-word-backend-254800.appspot.com/country/:country_code
```

By passing in the two-letter country code for a country, you can get data on the country relating to it's linguistic data, as well basic data on all the languages spoken in that country, as well as urls to graphs and maps related to the languages around that country

Path Variables

country_code	IN
--------------	----

Testing Tools

Mocha: For front-end testing, we used Mocha, which is a feature-rich JavaScript test framework running on Node.js in the browser. We chose Mocha because it provides functionality for testing both synchronous and asynchronous code with a very simple interface in the browser. It is also positive that it is open source and free.

Selenium: For more front-end testing, we used Selenium, an automated testing suite for web applications. This was paired with Mocha to create a comprehensive testing framework for our website. We chose Selenium because we needed to automate our web application for testing purposes, and it seemed to be the best choice that we could find. In addition, it is open-source, which we support.

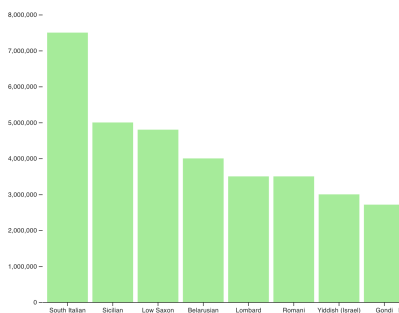
User Stories

The feedback that we got from our users will be very helpful in improving the functionality of our website. First, they suggested that links to new tabs be opened in a new tab. This is a good idea because the browsing experience of the user on our website will not be interrupted when they attempt to check out a charity. It also allows users to be able to multitask and look at the websites of multiple charities at the same time. Second, the customers suggested that the about page cards not be clickable. This was an oversight on our part, as the cards on the about page do not actually link to anything. A third request was that the country's flag is visible on the country page rather than the graphs. This may be more visually appealing for anyone that looks at the pages, and they can click on the 'Learn More' button to see the graph if they so desire. Fourth, the about page should contain a picture and a short blurb about each of the developers. This will create a more personal feel for the site. And last, it was suggested that we change the favicon so that the customers can quickly tell which tabs are open to our site if they have a lot open on their browser. Hosting We initially discussed many different hosting solutions, including Amazon Web Services, Google Cloud Platform, and Heroku, but ultimately settled on using Google Cloud Platform. We are hosting our backend Flask server and frontend React app separately.

Visualizations

For this phase, we created 3 different visualizations using the d3.js library, with one for each of our different models. For Languages, we visualized the most spoken languages on a bar graph. For Countries, we visualized the longitude of the countries we had in our dataset, also on a bar graph. Finally, for Charities, we created a bar graph for the income amounts of different charities in our database.

A visualization for some of the more commonly spoken languages in



A visualization for the income amounts of some of the charities in our dataset.

